Looking Back – and Forward: Reflecting on 25 Years in AI

September 17, 2020

By Nils Lenke

It's been 25 years since I completed my PhD in Natural Language Generation and joined this company (yes, Cerence is less than a year old, but we have a lengthy history!). In October 1984, I started studying something called "Linguistic Data Processing," so you could say that for the last 25 years – or even 36 years – I have worked on what we now call "Conversational AI." With reaching this milestone in my career, I've recently been reflecting on what has changed, what has stayed the same, and finally, what may be ahead of us when it comes to AI.

Firstly, there's been a shift in form factors. I spent the first several years of my professional career building "dialog systems" for the telephone (think "talking timetables" for trains). This was an advance over my university years, where most research in AI was on written text, even for dialog systems (the idea of a dialog on the basis of written text would come back much, much later as what we now know as "chatbots"). Of course, we're talking fixed line phones; the mobile phone didn't really come into mass usage before the late '90s (fun fact: I got my first mobile phone in 1997, and I still have the same number!). The smartphone made an appearance in 2007, and the smart speaker another decade later. So yes, form factors clearly have changed. On the other hand, my colleagues working on our embedded solutions started building speech recognition systems into cars in the '90s as well. Granted, initially it was just phone number entry, but still, the benefit was as clear then as it is now: when your hands and eyes are busy, voice comes in handy. That hasn't changed a bit.

The methods and algorithms have also changed, but that's not as clear as it might seem. Granted, in the 1980s and 1990s, "AI" mostly meant symbolic AI, trying to capture knowledge and reasoning in the form of symbols, logic and rules. But there was a niche even then where people were looking into artificial neural networks, although they could not really get them to work on meaningful problems. All Automatic Speech Recognition (ASR) was done with Machine Learning (ML) from the very beginning, but using something called "Hidden Markov Models" or HMMs, which are no longer the norm today. My colleagues in 1995 didn't even know – or would have accepted – that they were practicing AI; they instead described their work as "signal processing," and all of them had a degree in Electrical Engineering. Their bible was the Shannon & Weaver model of 1949, and "maximum likelihood" and "Bayes' law" ruled the world. I was hired as a software engineer, and on my first day my boss *literally* said to me, "Nils, we have not hired you because you have a PhD in AI, but although you have that degree." Then, around 2010, something funny happened. People found out how to get Deep Neural Networks (Neural Networks with more than one hidden layer) to work for image classification and then for speech recognition and how to use GPUs (invented for gaming consoles) to get the model trainings done in days rather than weeks or months. In 2012 we (as Nuance back then) had the world's first commercial ASR system in the cloud based on DNNs up and running, as well as the first connected cars, i.e. cars with a connection to servers in the cloud, that made use of it. Then, the general public got wind of the new topic and journalists picked it up. Cleary, from "artificial neural networks" we were constructing "artificial brains," and those clearly had to do "artificial intelligence," right? So, in a kind of ironic semantic shift, AI

became equal to Deep Learning (DL) or ML using DNNs. That never would have happened to the boring HMMs of old!

Today, we're already seeing the next shift in the approach to Conversational AI. The "traditional" approach of using DL was to apply the algorithms to a bunch of data specific to the task at hand. In Computer Vision (CV), things were always a bit different. The approach there was to train the system on a lot of generic data to learn some basic things (recognizing edges and basic shapes, etc.) and then train for specific tasks on top of that with less specific data. This approach has started to come to Conversational AI as well, first with word embeddings, then with BERT, and now with GPT-2 and GPT-3. Speaking of the latter, one word of caution. While the results of these systems are truly remarkable and impressive, one theme has also been constant over the last several decades: we tend to overestimate the "intelligence" in AI systems, all the way since Weizenbaum's ELIZA. Because of the anthropomorphism effect, as Naess called it, we automatically project human-like qualities into systems that speak or produce texts – we tend to assume there is true intelligence behind speech and text, even if there isn't. As advanced as they are, BERT and GPT-2 still make really "stupid" mistakes, showing that what they do is still pattern matching (albeit very advanced pattern matching).

Speaking of humans and AI, what has clearly changed is the attitude towards AI, especially Conversational AI. Up until about five years or so ago, we still had to explain why you should add ASR and NLU to a device. With the advent of assistants in smartphones and smart speakers, that has fundamentally changed. Today it is normal and expected that things can speak – and a disappointment if they can't. What hasn't changed, though, are some of the questions being asked, for example: "Can your system cope with accents? Particularly for my language X, which is so special." (Answer: yes, it can, because we include all accents and dialects in the training data, and no, your language is not "special" – all languages have that!).

Finally, what has not changed is the "domain" of Conversational AI, namely Human Speech & Language itself. Humans, ever since they turned into homo sapiens some 300,000 years ago, have spoken languages as complex as ours today. True, specific languages change (try reading some medieval English), but the nature and structure of the human language capability as such hasn't changed since then. Even when describing features of modern AI assistants, we use terms like anaphora (referring back to an entity mentioned earlier in the dialog) or ellipsis (leaving out words that can easily be inferred from the context), terms that were coined in ancient Greece more than 2,000 years ago.